東北大学機関リポジトリ
TOUR
Tohoku University Repository

日本語話し言葉を対象とした連続音声認識システムの高精度化に関する研究

著者	加藤 正治
号	54
学位授与機関	Tohoku University
学位授与番号	工博第4262号
URL	http://hdl.handle.net/10097/61612

	かとう　まさはる
氏名	加藤　正治
授与学位	博士（工学）
学位授与年月日	平成２２年３月２５日
学位授与の根拠法規	学位規則第４条第１項
研究科，専攻の名称	東北大学大学院工学研究科（博士課程）電気・通信工学専攻
学位論文題目	日本語話し言葉音声を対象とした連続音声認識システムの高精度化に関する研究
指導教員	東北大学准教授　伊藤　彰則
論文審査委員	主査　東北大学教授　牧野　正三　東北大学教授　鈴木　昭一 東北大学教授　安達　文幸　東北大学准教授　伊藤　彰則

論文内容要旨

音声は人と人とのコミュニケーションにおいて最も基本的かつ自然な手段である．

音声認識はコンピュータの可能性を具現化する好例として，古くから研究開発が行われてきている．現在求められていることは，自然発話を用いて情報検索できることや，インターネット上の至る所に存在する音声コンテンツに自由にアクセスできることであろう．

音声認識とは連続音声を単語列に変換する問題である．キーボードの代わりに音声を用いるデクテーションシステムのようにコンピュータに入力することを前提とする場合においては90%から95%程度の高い認識性能が得られ実用化ができている．しかしながら，講演や会議などの録音された音声を対象とする場合は，80%程度の認識性能にとどまっているのが実情である．デクテーションよりも難しい問題となる理由は，人間を相手とする発話が，音響的に不明瞭で，かつ，言語的に否定形になることが挙げられる．話し言葉を対象とする音声認識では，音響的・言語的な変動が大きくなり十分な認識性能が得られていない．

我々が日常生活において通常用いている自発性の高い自然な発話の「話し言葉」を自動認識する技術はこれからの情報化社会で必要不可欠である．認識システムの高精度化による性能改善が望まれている．

本論文は，音響モデル・単語辞書・言語モデルをそれぞれ高精度化することで認識性能を改善することを研究目的とする．

具体的な評価対象として，話し言葉の研究用のデータベース『日本語話し言葉コーパス(CSJ: Corpus of Spontaneous Japanese)』に含まれる講演音声を用いる．講演音声は独話(monologue)であり，一人が他の人に一方的に話すスタイルをとる．これは，最も基本的な話し言葉である．ここで構築された技術は，他の話し言葉（対話や会話など）にも応用が可能である．

講演音声認識は自発的な発話であることから認識が困難である．多くの場合，メモ程度参照があったとしても口頭原稿を読み上げるスタイルではない．また，話し手はアナウンサの様に発話に対する専門的な訓練を受けているわけではないので，発声も流暢ではなく認識はとても困難なものになる．

話し言葉では，書き言葉を模した発話や読み上げ音声と比べて発話速度の変動が大きい．特に発話が速くなると

調音結合の度合いが強くなる．より広い範囲での調音結合を考慮したモデルが必要である．現在，音声認識で広く使われている，triphone 音響モデルは前後 1 つの音素環境を考慮するがこれでは不十分である．これに対して，前後 2 つの音素環境を考慮する quinphone モデルが提案されている．音素環境依存のモデルは状態を共有することでモデルのサイズを制限し安定した認識精度が得られるように構築される．最適な状態数は，調音結合の度合いを考慮して設計することになる．一般に，詳細な（状態数の多い）モデルは強い調音結合に対応できるが，統計的に不安定になり推定精度が悪くなる．調音結合の弱い部分では粗い（状態数の少ない）モデルの方が推定精度は良い．

ここで問題となるのは，話速が話者間のみでなく，話者内，さらに，一つの発話内でも大きく変動することである．調音結合の強さもそれに伴って大きく変化する．したがって，最適なモデルを決めることができず，モデルの性能を安定して引き出すことができない．大きく変動する話速に対して頑健な評価ができるシステムの設計が必要となる．

話し言葉では，音素や音節単位で発音が変形することが観測されている．このため，調音結合として扱うには影響している範囲が広く，音響モデルだけでは対応できない．これらのエントリーを扱うためには単語辞書で発音変形をモデル化することが必要になる．さらに，複数の単語（形態素）にわたる大きな変形も起こっている．単語（形態素）の区切りを音響的な区切りとして扱うことができない．発音変形を形態素解析の段階で考慮しなければならない．

話し言葉には多様な話題が存在する．多くの話題が存在する理由として，

- 話し言葉の分野は様々である
- 発話内容は個人的なものを含めて多岐にわたる
- 語彙は話者が自由に選定している
- 発話スタイルは自由である．

一方で，従来まで研究されていた新聞記事の読み上げも自然言語を扱っている．また，ニュース音声も話し言葉の一つである．しかし，これらの音声には次のような制約がある．

- ニュースで扱うジャンルが決まっている
- 語彙が統制されている
- 発話スタイルも規定される

話し言葉の言語モデルは，書き言葉を基本とする音声認識よりも多様な話題を扱える必要がある．

話題を考慮する手法の一つに散在的意味解析（PLSA: Probabilistic Latent Semantic Analysis）と N-gram を組み合わせた研究がある．PLSA は潜在変数として topic（PLSA でいうところの話題）を用いて，一般的な意味での話題に適応している．多様な話題を取り扱うためには topic 数を増加させる必要がある．モデルの学習と適用には膨大な計算量と記憶容量が必要なであり，それらは topci 数に比例する．この問題のため大規模なモデルを学習することがこれまで不可能であった．話し言葉の多様な話題に適応するためには「大規模は PLSA の学習・適用法の開発」が不可欠である．

本研究では，話し言葉音声認識の高精度化に対する問題点として「**多様性**」と「**変動**」に着目した．

話し言葉音声認識において，次の問題に対処するために「**多数・大規模なモデルの利用と統合**」を提案する．

- 話速の変動による調音結合の影響のばらつき
- 発音変形による単語の発音の変動
- 多様な話題への言語モデルの対処法の必要性

本論文は5章で構成される．第1章は序論であり，話し言葉の音声認識の問題点を述べる．

第 2 章では，「音響モデルの高精度化」について，多数の音響モデルを併用し，それらを統合する手法を提案する．第 3 章では「単語辞書の高精度化」について，多様は発音変形のモデル化と言語モデルへの統合を提案する．第 4 章では，「言語モデルの高精度化」について，PLSA 言語モデルと N-gram 言語モデルを用いる手法において，これまで計算量の問題で実現不可能であった多様な話題を扱える言語モデルを構築する手法を提案する．第 5 章で結論を述べる．

以下に，本研究での成果を述べる．

第 2 章では，音響モデルの高精度化を目的として多数の quinphone 音響モデルを作成しスコアを統合する手法を提案した．話し言葉では，発話速度の変動が激しいために，調音結合の強さが変わる．話し言葉の音声の話速は，話者間だけでなく同一話者，さらに，同一発話内でも大きく変動する．発話速度の大きな変化のため最適なモデルを選択することはできない．

提案法は，多数の状態数の異なる quiphone 音響モデルを用いることで，モデルの「詳細さ」に変化をもたせ，変動する話速に対する頑健性を確保している．誤認識が起こる一つの理由として，正解単語のスコアに対して一部のモデルは低いスコアを出していることが挙げられる．しかし，多くのモデルは良いスコアを出すことが予測される．一方で不正解の単語は多くのモデルで低いスコアになっている．スコアを平均することにより，正解単語に対して低いスコアが与えられる危険が減るために性能が向上したと考えられる．

また，提案法では認識のときに多数のモデルから最適なモデルを選択する必要がない．多くの研究での最適なモデルの選択は，実験結果に基づいて事後的に行われている．また，選択基準そのものが問題となっている．モデルの選択が不要であることも利点の一つである．提案法は，話速の変動による調音結合の影響の変化を吸収することができた．従来法と比較して，10.7% の認識性能改善が得られた．

第 3 章では，話し言葉における多様な発声現象を捉えるための「発音変形依存モデル」を提案した．主に不明瞭な発話を中心とした発音変形の問題について検討した．話し言葉では，音素の置換や脱落といった調音結合では取り扱うことのできない大きな変形が観察されている．さらに，発音変形は単語の中だけでなく，複数の単語に渡ることもある．

発音変形はランダムに起こるわけではなく，ある種の言語的偏りがあることが想定される．そこで，実際の発話情報を，音響・言語情報に正確にモデル化することを試みた．複数の形態素に渡る変形は，これらの単語をまとめて新しいエントリーとすることを提案した．形態素解析と発音の割り当てを同時に考慮することで，実際の話し言葉に含まれる様々な発音情報をモデリング化することができた．言語モデルの発音変形への対処は効果が高く，提

案したモデル化は従来法と比較して 26.5% の認識性能の改善を得られた.

第 4 章では, 話題を考慮した PLSA 言語モデルによる高精度化について述べた. 話し言葉には多様な話題が含まれている. そのため, 単純な N-gram では大局的な単語の依存関係を考慮できない. これに対して, PLSA 言語モデルが提案されているが, 従来までの研究ではモデルサイズが十分ではなく, 話し言葉の話題を扱いきれていない. 話し言葉は従来考えられてきたよりも多くの話題から成り立っていると予想される.

多様な話題をモデル化するためには, PLSA の話題サイズを大きくする必要がある. しかし, PLSA の計算量は話題サイズに比例して増加する. PLSA の学習は, 従来から計算量が多いことが指摘されていて, 話題サイズを大きくすることは困難であり, 実現可能な学習と適応アルゴリズムの構築が不可欠であった. この問題に対して本研究では, 並列処理による高速学習アルゴリズムを提案した. 並列処理の実装には LAM/MPI ライブラリを用いている. また, 適応においての計算量を低減する手法についても検討し, unigram rescaling において N-gram の backoff を考慮した高速化手順を示した. これまでより, 多数の話題をもつ PLSA モデルを作成することが可能となり, 従来法と比べて 5.9% の認識性能改善を得ることができた. 言語モデル単独でこれだけの認識性能を改善していることの意義は大きい.

各章の手法を統合した実験結果を図 1 に示す. 単語誤り率は 17.42% を得ている. 第 2 章の実験で単独のモデルを用いた場合の最良値の 19.22% と比較すると, 9.4% と大きな改善率を得ることができた. さらに, 音響適応によく用いられている MLLR 適応を併用した場合, 単語誤り率 14.92% の高い性能を得ることができた.

本研究は, 音韻・発音・語彙が多様な話し言葉に対し, 多様かつ大規模なモデルを利用統合することで, 認識の高精度化を達成した. このことの意義は大きい.

評価は CSJ の講演音声で行われているが, より自発性の高い, 対話音声や会議音声などへの応用も可能である.

本研究での話し言葉の音声認識を高精度化は, 音声言語の研究の基盤となる手法である. 音声情報の活用の利便性を向上させることができ社会的意義も大きい.

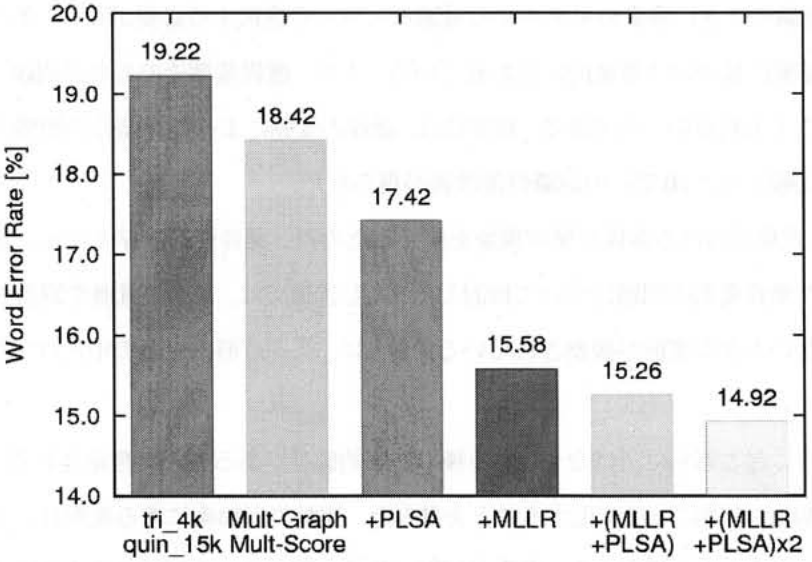

図 1 単語グラフの構造・スコア統合と PLSA・MLLR 適応

論文審査結果の要旨

丁寧に発話した日本語連続音声を対象とした音声認識システムが既に商用化されているが，講演や会議などで話される音声（話し言葉）では音響的，言語的変動が大きく，高い精度が得られていない。話し言葉で現れる音響的，言語的変動を克服する方法を開発することは，次世代の講演自動書き起こしシステムや自動要約システムの構築を進める上で極めて重要である。著者は，多数の音響モデル，複数の辞書項目表現，多数の話題モデルを認識システムに備えることによって，音響的変動と言語的変動を吸収する方法を提案し，日本語話し言葉コーパスを用いてその有効性の検証を行った。本論文はその研究成果をまとめたもので，全編５章からなる。

第１章は序論であり，本研究の背景及び目的を述べている。

第２章では，多数の音響モデルを併用して，話速の変動に頑健な音声認識を行う方法を提案した。従来は，話速が速い音声における音素の音響的変化を表現するため，前後２個の音素環境を考慮する音素モデルである quinphone が用いられていた。しかし，話速の変動が大きい音声に対しては，どの程度詳細な quinphone を利用するのが良いかを最適に決定することは難しかった。提案法では，様々な詳細度を持つ quinphone を複数併用した上で，それらから計算されるスコアを統合して音声認識結果を推定する方法を提案し，単語誤り率を 4.3%改善した。これは，変動が大きい話し言葉を認識するための重要な成果である。

第３章では，話し言葉における発音変形に頑健な認識手法を提案した。話し言葉では，単語における音素や音節の置換や脱落が発生し，認識性能低下の原因となる。従来は，単語ごとに発音の変形を辞書項目として登録する方法が取られていたが，この方法では複数の単語にわたる発音変形を表現することができない。これに対して提案法では，発音変形の有無によって単語の定義を使い分けると同時に，発音と表記との組み合わせを新たに単語の定義とする。提案法によって，単語誤り率が 26.5%削減された。これは，話し言葉における発音変形に頑健な音声認識手法として重要な成果であり，高く評価できる。

第４章では，確率的潜在意味解析（PLSA）を用いた言語モデルを高速に学習するアルゴリズムを提案した。PLSA は言語モデルを話題に適応させる手法として知られていたが，計算量が多いため，これまで大規模なモデルを構築することは困難であった。これに対して提案法では，並列化アルゴリズムを用いて計算時間を短縮するとともに，確率計算の計算量を低減するアルゴリズムによって計算時間を 6000 倍以上高速化した。また，ここで提案した方法を用いて学習した大規模 PLSA モデルを用いることにより，従来法と比べて単語誤り率が 5.9%改善された。これは，多数の話題に依存した言語モデルの有効性を示した重要な成果であり，高く評価できる。

第５章は結論である。

以上要するに本論文は，話し言葉音声の音響的・言語的変動に対して頑健な音声認識手法を提案したものであり，音声認識工学および電気・通信工学の発展に寄与するところが少なくない。

よって，本論文は博士(工学)の学位論文として合格と認める。